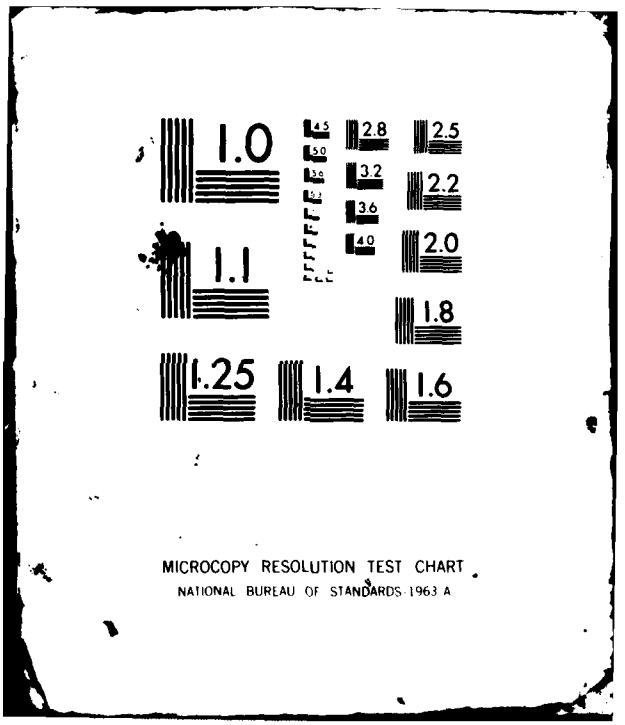
AD-A111 532 MASSACHUSETTS UNIV AMHERST DEPT OF MATHEMATICS AND S--ETC F/G 12/1
NONPARAMETRIC TESTS OF INDEPENDENCE AND GOODNESS-OF-FIT FOR CEN--ETC(U)
DEC 81 R M KORWAR AFOSR-80-0219

UNCLASSIFIED AFOSR-TR-82-0072 NL

END
DATE FILMED
3 82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963 A

12

NONPARAMETRIC TESTS OF INDEPENDENCE AND GOODNESS-OF-FIT FOR CENSORED DATA[1]

by

RAMESH M. KORWAR

AFOSR Final Technical Report

December, 1981

The University of Massachusetts
Department of Mathematics and Statistics
Amherst, Massachusetts

[1]Research sponsored by the Air Force Office of Scientific Research, AFSC,USAF under Grant AFOSR-80-0219 and modified AFOSR-80-0219A.

ABSTRACT

The work accomplished is represented by four Tech Reports already issued and the development of three tests of goodness-of-fit for censored data reported herein. All the Tech Reports are submitted for publication. Two of the Tests are developed using a result due to Moses (J. Amer. Statist. Assoc. 59, (1964), 645-51) for uncensored data and its modification for the censored data. The other is an extension of the empty cell test to the censored case.

1. *Introduction.*

The accomplishments are represented by the following Technical Reports (listed in chronological order) written and issued from time to time, and the work on three tests of goodness-of-fit for censored data reported herein below:

[1] Korwar, R.M. (1980). A characterization of a Polya-Eggenberger and other discrete distributions by record values.

[2] Korwar, R.M. (1981). A characterization of the Waring distribution.

[3] Korwar, R.M., and Naik, D.N. (1981). Testing for equality of means with additional data on one variable: a likelihood ratio test and a Monte Carlo study.

[4] Korwar, R.M. (1981). On characterizations of the power-function and discrete uniform distributions through a model of over-reported claims.

2. A Brief Description of the Work Reported in [1]-[4].

In [1] above, a class of Polya-Eggenberger distributions is characterized by record values. The Polya-Eggenberger distribution is one of the truly "contagious distributions" found very useful in applied work. Specifically, let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed discrete random variables. Define the seq. $\{N(n)\}$ by $N(1) = 1$, $N(n) = \min\{j | j > N(n-1), X_j > X_{N(n-1)}\}$, $n = 2,3,\ldots$. Let $R_n = X_{N(n)}$. Then $\{R_n\}$ is the sequence of record values. By convention $R_1 = X_1$. Assume $E(X_1)$ exists and is finite. Here characterization of a Polya-Eggenberger and other discrete distributions, including the geometric is made by the linearity of

of regression of $R_2 - R_1$ on R_1.

This paper is submitted to Sankhyā for publication. An abstract has appeared in the Bulletin of the Institute of Mathematical Statistics (IMS Bulletin 10, #2(1981), 64, #8/t-33).

In [2] above, a characterization of the Waring distribution is made by the identity of distributions. The Yule distribution, which is sometimes used as a distribution of word frequencies in applied work, is a special case of the Waring distribution. It is characterized by the following property: For a positive integer-valued random variable X, $P(X=r) = p_r$, $r = 1,2,\ldots$, and with a finite mean μ define two new random variables Y and Z by

$$P(Y=r) = q_r = \sum_{k=r+1}^{\infty} p_k + ap_r)/(\mu+a), r = 0,1,\ldots$$

$$P(Z=r) = q'_r = (r+b)p_r/(\mu+b), \quad r = 1,2,\ldots$$

where $a \geq 0$ and b are constants with $b - a + 1 > 0$. Then Z and Y truncated at 0 have the same distribution if and only if X has a Waring distribution of the form

$$P(X=r) = (\lambda-c)c^{[r-1]}/\lambda^{[r]}, \quad r = 1,2,\ldots,$$

where $\lambda - c > 1$, $c > 0$; and $c^{[r]} = c(c+1)\ldots(c+r-1)$, $r = 1,2,\ldots$, $c^{[0]} = 1$.

This manuscript has been submitted for publication to Sankhyā. An abstract has appeared in the Bulletin of the Institute of Mathematical Statistics (IMS Bulletin 10, #4(1981), 158, #81t-70).

In [3] above, a likelihood test is derived for testing the equality of means of a bivariate normal distribution with equal variances when additional data on one variable are available. The situation can also be viewed as if some observations on one variable are missing. A Monte Carlo study is conducted to study the power and level of significance attained in an attempt at comparing several tests available in the literature along with the proposed test. As a result of the study an indication is made of the preferred test for each combination of the correlation coefficient and difference of means.

This was submitted to the Journal of American Statistical Association and a revision is underway. The revisioned version will be resubmitted to the above journal or somewhere else. An abstract is submitted and will appear in the Bulletin of the Institute of Mathematical Statistics. This research is a natural counterpart to Dahiya and Korwar (1980).

In [4] above, using a model for over-reported claims (such as insurance claims for fire damage to property, etc.) some characterizations of useful distributions in statistics are made. Using this model which assumes overreporting the power-function and discrete uniform distributions are characterized as follows: (1) The distribution of observed claims suitably truncated on the right coincides with the true distribution if and only if the distribution is of the power-function form and (2) a variable having a linear regression on the true claims has a linear regression, with suitable slope and intercept, on the reported claims if and only if the distribution is of the power-function

form. Similar results are obtained for the discrete uniform distribution.

3. Two Tests of Goodness-of-fit for Censored Data Based on a Result of Moses.

Suppose $Y_1^0,\ldots,Y_n^0$ is a sample of size n from a continuous distribution G. Due to random censoring on the right we do not observe the Y_i^0's but

$$Y_i = \min(Y_i^0, U_i),\ i = 1,\ldots,n \tag{3.1}$$

where $U_1,\ldots,U_n$ are independent random variables (r.v.), called censoring r.v.'s, with a continuous distribution function H. Assume that Y_i^0's and U_i's are mutually independent. We also observe

$$\delta_i = \begin{cases} 1, \text{ if } Y_i^0 \le U_i \\ 0, \text{ otherwise} \end{cases} \tag{3.2}$$

for $i = 1,\ldots,n$. The problem is, given the censored data

$$\{Y_i, \delta_i,\ i = 1,\ldots,n\} \tag{3.3}$$

to test whether Y_i^0's could have come from a specified distribution F. That is we would like to test

$$H_o: G(y) = F(y) \quad \text{all real } y \tag{3.4}$$

against

(3.5) H_1: $G(y) \neq F(y)$ for some real y.

We derive two tests for testing H_0 by making use of Moses'(1964) one sample limits of some two-sample rank tests. Let $X_1,\ldots,X_m$ be a sample of size m from F and let $Y_1,\ldots,Y_n$ be an independent sample of size n from G. Both the distribution functions F and G are assumed unknown. Then Moses showed that the limit, as $m \to \infty$, of Lehmann's most powerful test of

$$H_0: F(x) = G(x) \quad \text{all real } x$$

against

$$H_1: G(x) = [F(x)]^k, \quad k > 1$$

is to reject H_0 for large values of

(3.6) $$\sum_{j=1}^{n} \ln F(Y_j) .$$

Note that now F becomes known since $m \to \infty$ and we have an infinite sample from F. Similarly he shows the limit, as $m \to \infty$, of the Wilcoxon two-sample test of

$$H_0: F(x) = G(x) \quad \text{all real } x$$

against

$$H_1: F(x) > G(x) \quad \text{all real } x$$

is to reject H_0 for large values of

(3.7) $$\sum_{j=1}^{n} F(Y_j).$$

Now back to H_0(3.4). We cannot directly use (3.7) with censored data. Because of censoring some of the $F(Y_j^0)$ cannot be computed. We replace (3.7) by its conditional expectation given G = F and the data (3.3). Thus our test statistic will be

$$T_n = E(\Sigma F(Y_j^0) | F, Y_i, \delta_i, i=1,\ldots,n).$$

But

$$E(F(Y_j^0) | F, Y_j, \delta_j = 1) = F(Y_j),$$

and

$$E(F(Y_j^0) | F, Y_j, \delta_j = 0) = \int_{Y_j}^{\infty} F(y)dF(y) / \int_{Y_j}^{\infty} dF(y)$$
$$= \frac{1}{2}\{1+F(Y_j)\}.$$

Thus, we take as our test statistic

$$T_n = \sum_{j=1}^{n} \delta_j F(Y_j) + \frac{1}{2} \sum_{j=1}^{n} (1-\delta_j)\{1+F(Y_j)\} \tag{3.8}$$

$$= \sum_{j=1}^{n} V_j,$$

where

$$V_j = \frac{1}{2}\{(1+\delta_j)F(Y_j) + (1-\delta_j)\} \tag{3.9}$$

In the following theorem we prove the asymptotic normality of T_n.

Theorem 3.1: The statistic T_n is asymptotically normal with asymptotic mean and variance n_μ and no σ^2 where

$$(3.10)\quad 2\mu = 2E(V_1)$$

$$= 1 - \int_0^\infty G(u)dH(u) + \int_0^\infty F(y)dG(y) + \int_0^\infty \left\{\int_0^u F(y)dG(u)\right\}dH(u)\,,$$

$$(3.11)\quad \sigma^2 = \mathrm{Var}(V_1)$$

and

$$(3.12)\quad 4E(V_1^2) = 1 - \int_0^\infty G(u)dH(u) + \int_0^\infty F^2(y)dG(y) + 3\int_0^\infty \left\{\int_0^u F^2(y)dG(y)\right\}dH(u)$$

$$2\int_0^\infty \left\{\int_0^u F(y)dG(y)\right\}dH(u)\,.$$

Proof: The theorem follows from the central limit theorem and the fact that V_j's are independent and identically distributed bounded random variables with common mean and variance given by (3.10-(3.12).

Note that since $G = F$ under H_0 the asymptotic null mean μ_0 and variance σ_0^2 are given by (3.10)-(3.12) where we replace G by F. The censoring distribution H appearing in (3.10)-(3.12) is generally unknown and must be estimated from the data.

The estimation of $\bar{H}(u) = 1 - H(u)$ from the data by the method of Kaplan and Meier (1958) is completely analogous to the estimation of $\bar{G}(y) = 1 - G(y)$ from the data and using the same method, except for the fact that $(1-\delta_j)$'s now play the role of δ_j's before. Let $Y_{(1)} < \ldots < Y_{(n)}$ be the ordered Y_j's and let $\epsilon_j = 1 - \delta_{[j]}$, where

$\delta_{[j]}$ is the δ that goes with $Y_{[j]}, j = 1,\ldots,n$. Then the Kaplan-Meier (K-L) estimator $\hat{\bar{H}}(u)$ of $\bar{H}(u)$ is given by

$$(3.13) \qquad \hat{\bar{H}}(u) = \prod_{j=1}^{k-1} \{(n-j)/(n-j+1)\}^{\epsilon_j}, \quad u \in (Y_{(k-1)}, Y_k],$$

and $\bar{H}(u) = 0$ for $u > Y_{(n)}$. Thus consistent estimators $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ of μ_0 and σ_0^2 respectively can be obtained from μ_0 and σ_0^2 by replacing H by $\hat{H}(3.13)$ appearing in their expressions. The consistency of the resulting estimators of μ_0 and σ_0^2 follows from the weak convergence of the K-L estimator. Combining Theorem 3.1 and the above we have

$$(3.14) \qquad Z_n = (T_n - n\hat{\mu}_0)/\sqrt{n}\hat{\sigma}_0 \xrightarrow{L} N(0,1), \quad n \to \infty.$$

Finally, to test H_0 against H_1 at level α, we reject H_0 if $|Z_n| > Z_{\alpha/2}$ and accept otherwise, where $Z_{\alpha/2}$ is the $(1-\alpha/2)100$-th percantile for the standard normal distribution.

A similar test can be constructed using (3.6) and the same technique of replacing the test statistic for the uncensored case by its conditional expectation given $G = F$ and the data for the censored data case. The resulting test statistic will have asymptotic normality since the test statistic again is going to be a sum of independent and identically distributed random variables.

Hollander and Proschan (1979) use the same idea due to Moses and come up with a test different from our tests.

4. An empty cell test.

In this section we derive an empty cell test for censored data. We use the notation developed in Section 3. Using the hypothesized continuous distribution function F, choose points $x_0 = -\infty < x_1 < \ldots < x_{N-1} < x_N = \infty$ such that $F(x_k) - F(x_{k-1}) = 1/N$, for $k = 1,\ldots,N$, where N is a specified positive integer. In the uncensored case the test statistic used is

(4.1) $\mu_0(n,N)$ = # of intervals $(x_{k-1},x_k]$ containing no observation Y_j^0's

The test is to reject H_0(3.4) if $\mu_0 \leq C$, where C is chosen to have a α level test. The empty cell test is attractive because of its simplicity. An excellent reference on the subject is the recent book by Kolchin et al (1978).

Because of right censoring not all the Y_j^0's are observed. Hence μ_0(4.1) cannot in general be computed. We replace μ_0 by its conditional expectation given $G = F$ and the censored data (3.3) and use the resulting random variable as the test statistic. Let

(4.2) $\mu_d(n,N)$ = # of apparent empty cells,

(4.3) $C_i = (x_{k_i-1}, x_{k_i}]$, $i = 1,\ldots,\mu_d$, ith apparent empty cell.

Then, it can be shown that

$$\text{of } \mu_0^*(n,N) = E\{\mu_0(n,N) \mid Y_j, \delta_j, j = 1,\ldots,n\}$$

$$= \sum_{i=1}^{\mu_d} \prod_{j=1}^{n} \left[1 - \frac{\min\left\{F(x_{k_i-1}), F(Y_j)\right\} - \min\left\{F(y_j), F(x_{k_i})\right\}}{F(Y_j)} \right]^{1-\delta_j}$$

The distribution theory, both small and large sample, μ_0^* is now being derived and will be reported later.

REFERENCES

Dahiya, R.C., and Korwar, R.M. (1980). Maximum likelihood estimates for a bivariate normal distribution with missing data. Ann. Statist. 8, 687-92.

Hollander, M., and Proschan, F. (1979). Testing to determine the underlying distribution using randomly censored data. Biometrics 35, 393-401.

Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53, 457-81.

Kolchin, V.F., Sevast'yanov, B.A., and Chistyanov, V.P. (1978). Random Allocations. Wiley, New York.

Moses, L.E. (1964). One sample limits of some two-sample tests. J. Amer. Statist. Assoc. 59, 645-51.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFOSR-TR- 82 -0072	2. GOVT ACCESSION NO.: AD-A111532
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Nonparametric Tests of Independence and Goodness -of-fit for Censored Data	5. TYPE OF REPORT & PERIOD COVERED: Tech Report, Final June 15,1980 to Oct. 14, 1981
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Ramesh M. Korwar	8. CONTRACT OR GRANT NUMBER(s): AFOSR-80-0219
9. PERFORMING ORGANIZATION NAME AND ADDRESS: The University of Massachusetts, Department of Mathematics & Statistics, Amherst, MA 01003	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research, Bolling Air Force Base, D.C. 20332	12. REPORT DATE: December 1981
	13. NUMBER OF PAGES: 11
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release: distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Censored data
Tests of goodness-of-fit
Empty cell test
Asymptotic normality

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The work accomplished is represented by four Tech Reports already issued and the development of three tests of goodness-of-fit for censored data reported herein. All the Tech Reports are submitted for publication. Two of the tests are developed using a result due to Moses (J. Amer. Statist. Assoc. 59 ,(1964), 645-51) for uncensored data and its modification for the censored data. The other is an extension of the empty cell test to the censored case.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


DAT
FILM